3574–3577 Nucleic Acids Research, 1994, Vol. 22, No. 17

OWL — a non-redundant composite protein sequence database

A.J.Bleasby, D.Akrigg[1]* and T.K.Attwood[2]
DRAL, Warrington, Cheshire WA4 4AD, [1]Departments of Biochemistry and Molecular Biology, University of Leeds, Leeds LS2 9JT and [2]University College London, London WC1E 6BT, UK

ABSTRACT

A comprehensive, non-redundant composite protein sequence database is described. The database, OWL, is an amalgam of data from six publicly-available primary sources, and is generated using strict redundancy criteria. The database is updated monthly and its size has increased almost eight-fold in thelast six years: the current version contains >76000 entries. For added flexibility, OWL is distributed with a tailor-made query language, together with a number of programs for database exploration, information retrieval andsequence analysis, which together form an integrated database and software resource for protein sequences.

INTRODUCTION

Protein and nucleic acid sequence databases are now established as essential tools for the molecular biologist. Computer analysis of database sequences facilitates the identification of functional and structural motifs, and results of such studies may be used to assist both experimental (e.g. mutagenesis) and theoretical (e.g. structure prediction) studies. Biological databases must therefore be up-to-date, accurate, well-annotated and compatible with the most efficient software for information retrieval, data manipulation and similarity searching. Difficulties in achieving these goals arise partly because of the recent explosion in the volume of available sequence data, but also because of the problems associated with the number of different database standards, concepts and structures used by the laboratories compiling the information.

Protein sequence databases are distributed as alphanumeric files containing sequential lists of entries: these consist of data fields that include both bibliographic and textual information, in addition to the sequence itself. The principal uses of sequence databases are for similarity searches, sequence alignment, pattern recognition and information retrieval. If analyses are to be as comprehensive and up-to-date as possible, the proliferation of different databases raises a number of problems: these include the need to search all the primary sources; the need to reformat databases and to manage the use of software designed for different database formats; and the occurrence of considerable redundancy between the different sources. This latter is particularly inefficient and wasteful, in terms of the unnecessary expenditure both of computer time in scanning large numbers of identical sequences, and user time in wading through substantial redundancy in search output.

There are a number of ways to tackle a number of these problems, but in practice few solutions attack them all. One approach is to design software thatis sufficiently flexible to communicate with a variety of database formats and will perform sequential searches on a list of user-specified databases (e.g. as in GCG [1]) — but this does not efficiently address the problem of redundancy (as a trivial example, a relevant search would find the 104 KD microneme rhoptry antigen in both SWISS-PROT, with accession number P15711, and PIR, with accession number A44945). Another strategy is to search a nucleic acid sequence database as its amino acid translation in all six reading frames, using a method such as TFASTA [2]. This has the advantage that such databases (e.g. EMBL [3] and GenBank [4]) tend to be relatively up-to-date, but causes problems with data accuracy because of the occurrence of artefacts through translation of introns and other non-codings sequences. Our own approach has been to create the OWL database, a non-redundant composite of the major publicly-available primary sources, including a translated nucleic acid sequence database [5]; with it we provide appropriate software for its interrogation and exploration [6]. The composite directly addresses the need for comprehensive, non-redundant, efficient searches, but does engender various practical problems,arising primarily from the difficulties in manipulating and storing such large amounts of data.

Primary databases have different standards of validation and annotation of submitted entries, so a good composite database should preferentially include entries from the best-validated and best-annotated source. The COMPO software, which is used to generate our composite database [5], allows the user to specify which source is of higher priority: at the time of compiling the first release, this was NBRF-PIR [7]. Various developments have now led us to alter the composition of OWL, both in terms of its constituent databases and the order in which they are incorporated into the composite.

DATABASE COMPOSITION

The original sources included in OWL were NBRF-PIR [7], SWISS-PROT [8], a GenBank translation retrieved from the feature tables [4,9], NBRF-NEW, NEWAT86 [10], PSD-

*To whom correspondence should be addressed

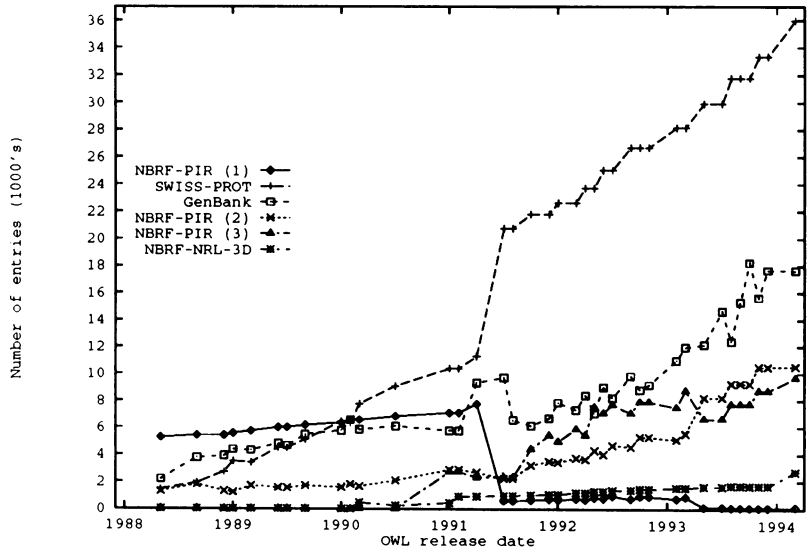

Figure 1. Graphs illustrating the contributions to the OWL database of its major sources from the earliest release to the current version. Contributions from SWISS-PROT can be seen to rise at a greater rate than all other sources, and the dramatic change in the shape of the graph at version 12.0 reflects the consequent replacement of PIR with SWISS-PROT as the highest priority source.

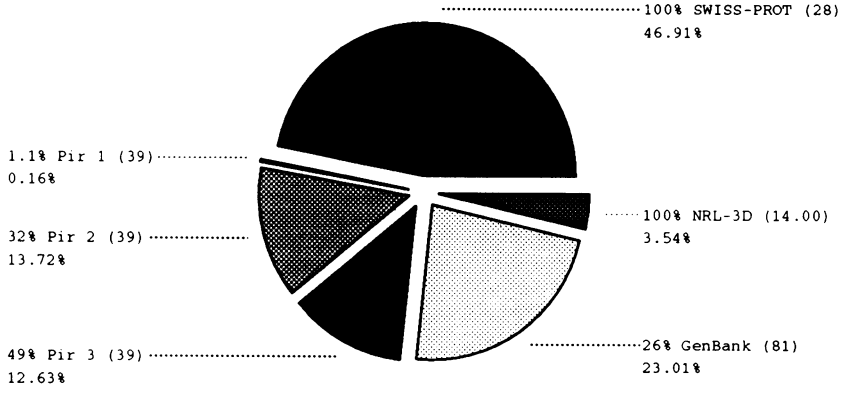

Figure 2. Pie chart illustrating the current composition of OWL (version 23.0).The chart shows the percentage of each source database (version number in parentheses) included in the composite, together with the percentage of OWL that this contribution represents. As the highest priority source, the whole of SWISS-PROT is included and accounts for almost 50% of OWL.

KYOTO [11], and the sequences contained in the Brookhaven protein structure databank [12]. Two of these sources, however, were not regularly maintained, and a point was reached at which it was felt that they no longer contained sufficient unique data to justify their retention in OWL. Other changes arose from (i) the distribution of three PIR databases, with different degrees of validation and commenting, where formerly there had been two [13], and (ii) the release of a new database of sequences derived from the structure databank, namely NRL-3D [14]. One additional source was thereby incorporated into OWL, while NRL-3D replaced the former Brookhaven extraction owing to better validation of sequence information.

Within OWL, each of the constituent databases is assigned a priority with respect to sequence validation and annotation; each is then compared against the highest priority source; redundant entries are eliminated, using strict criteria imposed by the COMPO suite; and the results are finally merged. Initially, PIR was assigned the highest priority, and all other databases were compared against it. Changes in the quality of individual databases, however, prompted a further change to OWL, with SWISS-PROT ultimately replacing PIR as the higher priority source. This was desirable because SWISS-PROT was emerging as the most comprehensive database, as shown in Figure 1, and had the most extensive and most structured annotation.

The current version of OWL (23.0) reflects the merging of six components, in the order SWISS-PROT, PIR 1, PIR 2, PIR 3, GenBank and NRL-3D. Figure 2 shows the relative contributions of each of these elements: SWISS-PROT is included in its entirety and all other sources are compared against it; only 1.1% of PIR 1 is retained, indicating the high degree of correspondence between the twodatabases; 32% of PIR 2 is added; 49% PIR 3; 26% of GenBank; and 100% NRL-3D. The complete inclusion of this latter database, we feel, is worth the small compromise to the non-redundant status of OWL for the value in knowing

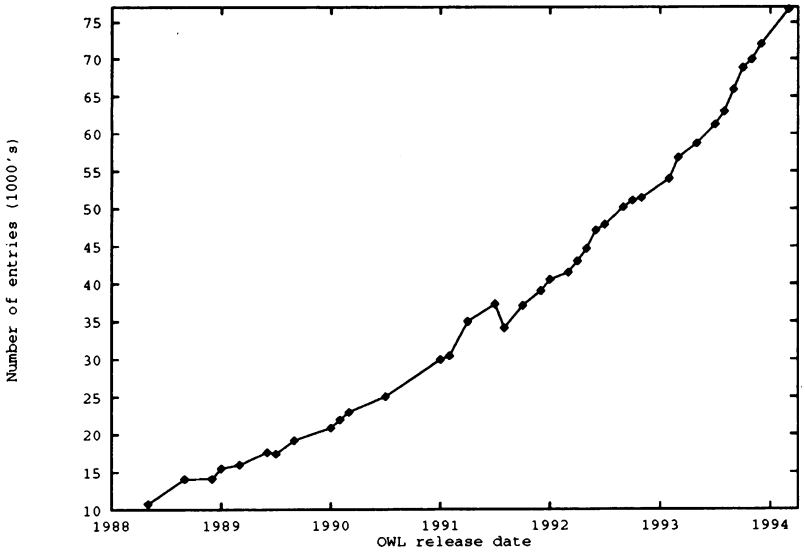

Figure 3. Graph showing the growth in the total number of entries in OWL from the earliest release to the current version. The small dip in mid-1991 results from the enhanced redundancy checking that was implemented on changing the priority of SWISS-PROT and PIR.

which are the sequences of known structure (it accounts for about 3% of the database). Note that entries in the composite retain the protein identificationcodes assigned by their particular sources, but those from NRL-3D are prefixed by NRL__ in order that sequences with known structures are immediately identifiable, and in cases where a GenBank sequence codes for more than one protein, the codes are suffixed with the corresponding number.

REDUNDANCY CRITERIA

OWL contains no sequences that are 100% identical, nor does it include sequences containing only 'trivial' differences. Sequence comparison is effected using contiguous segments of 30 residues, and sequences are included if there is more than one mismatching segment between a compared pair: a mismatching segment is defined as a segment that does not share 100% identity with an existing database sequence. A length criterion allows replacement of one sequence by another if the query sequence gives 100% identity but exceedsthe length of an existing entry. A complete description of the redundancy criteria is given in [5].

The OWL database therefore differs from other composites (e.g. NRDB [15] and MIPSX [13]) in that not only are identical sequences excluded but so also are sequences that only show minor differences. OWL is thus smaller than these resources, because it is more truly non-redundant. Database searches are consequently more efficient and resulting hitlists contain less noise.

DATABASE UPDATE AND GROWTH

OWL is released in major and minor versions. Major versions represent updates of the highest priority source, acquired from the international Internet ftp server (expasy.hcuge.ch). Minor versions are compiled whenever GenBank or PIR are released (e.g. from the ftp server ncbi.nlm.nih.gov), which tends to happen in a staggered manner, between releases of SWISS-PROT. This allows OWL to be built approximately every 4–6 weeks. The current version is 23.0, which contains 76,729 sequences, reflecting a near eight-fold increase in the size ofthe database in the last six years, as illustrated in Figure 3. This rate will undoubtedly increase once the human and other genome projects get fully underway. Note that the growth rate depicted here is virtually a linear rise in the number of sequences with time, rather than an exponential rise, as is often reported for primary source rates. This is because OWL sequence data reflect non-redundant growth.

APPLICATIONS

OWL has proved robust in wide and frequent use in the molecular biology community for numerous sequence similarity searches, sequence pattern analyses and for information retrieval. Specifically, it is the source database for the MOWSE peptide mass database [16], which provides peptide mass 'fingerprint' maps that allow unique and rapid identification of unknown sample proteins. MOWSE is available both to registered users of the UK SEQNET service and also via an email server (send an email message to mowse@dl.ac.uk containing the word 'help' in the message body for details).

OWL also provides the source sequence data for the PRINTS protein motif fingerprint database [17], which provides protein family signatures for fast diagnosis of newly-sequenced proteins. PRINTS is available via the SEQNET (s-ind2.dl.ac.uk), NCBI (ncbi.nlm.nih.gov) and EMBL (ftp.embl-heidelberg.de) anonymous-ftp servers, and is also distributed on the EMBL suite of CD-ROMs.

ASSOCIATED SOFTWARE

OWL and PRINTS are the central components of an integrated database and software resource that allows database exploration, pattern recognition and information retrieval [6]. The package

includes query languages for each of the databases; global and local similarity search facilities [18,19]; and a varietyof other programs for multiple sequence alignment [20] and pattern recognition [21]. This software can be accessed by UK academics via the SEQNET facility at Daresbury (for further details, contact the authors).

DATABASE DISTRIBUTION

OWL is generated in the form of a flat sequence file, together with a number ofindex files — these relate to the principal fields within individual database entries (i.e. codes, titles, text and sequence), which means that interrogationsoftware is fast, because queries can be directed to specific fields. The penalty for using an index system of this type is that the resulting files are very large, so the database occupies an increasingly large amount of disc space(currently 0.6Gb).

OWL is available via anonymous ftp from many sites, including s-ind2.dl.ac.uk(Europe/Africa/Asia), and ncbi.nlm.nih.gov (USA/Pacific rim). It is provided in two formats: PIR format allows system managers to access the database using theindex and query programs provided in the GCG package [1]; the FASTA [2] format file allows access via the BLAST [22] and FASTA suites. All common formats are therefore handled, allowing all software suites to access OWL on the user's local system.

Interactive access is available using a WorldWide Web server (URL:http://www.gdb.org/Dan/proteins/owl.html), which allows boolean operationsand cross-referencing to other common molecular biology databases. OWL can alsobe accessed via SEQNET using its fast free-text indexing system and query language (contact the authors for further details).

A BLAST server for OWL will shortly become available.

FUTURE DIRECTIONS

There are several further developments of OWL that we are addressing for future releases. First, we aim to achieve closer integration with the highest prioritysource, in order to make SWISS-PROT more comprensive and thus improve the extent of annotation in OWL. Second, whilst the composite database benefits from being extremely comprehensive and non-redundant, it clearly lags behind those for which nightly updates are possible. Most source databases now provideftp updates either nightly or weekly. To address this situation, we therefore intend to make an update database available (OWLET) between major and minor releases of OWL.

ACKNOWLEDGEMENTS

We are grateful to Dan Jacobsen for production of the OWL-WEB WorldWide Web server. We thank Michael Beck for preparing the figures. Funds for the originaldatabase project were provided by the SERC Protein Engineering Club. We are grateful to the University of Leeds for their continuing support of this work. TKA is a Royal Society University Research Fellow. This work benefitted from the use of the DRAL SEQNET facility.

REFERENCES

1. Devereux, J., Haeberli, P. and Smithies, O. (1984) Nucleic Acids Res., 12 (1), 387–395.
2. Pearson, W.R. and Lipman, D.J. (1988) Proc.Natl.Acad.Sci.USA., 85, 2444–2448.
3. Hamm, G.H. and Cameron, G.N. (1986) Nucleic Acids Res., 14, 5–10.
4. Burks, C., Fickett, J.W., Goad, W.B., Kanehisha, M., Lewitter, F.I., Rindone, W.P., Swindell, C.D., Tung, C-S. and Bilofsky, H.S. (1986) CABIOS, 1, 225–233.
5. Bleasby, A.J. and Wootton, J.C. (1990) Protein Engineering, 3 (3), 153–159.
6. Akrigg, D., Attwood, T.K, Bleasby, A.J., Findlay, J.B.C., Maughan, N.A., North, A.C.T., Parry-Smith, D.J., Perkins, D.N. and Wootton, J.C. (1992) CABIOS, 8 (3), 295–296.
7. George, D.G., Barker, W.C. and Hunt, L.T. (1986) Nucleic Acids Res., 14, 11–15.
8. Bairoch, A. and Boeckmann, B. (1993) Nucleic Acids Res., 21 (13), 3093–3096.
9. Fickett, J.W. (1986) Trends Biochem.Sci., 11, 190.
10. Doolittle, R.F. (1981) Science, 214, 149–159.
11. Kanehisa, M., Institute for Chemical Research, Kyoto University, Uji, Kyoto 611, Japan.
12. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, D.F., Brice, M.D.,Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977) J.Mol.Biol.,112, 535–542.
13. Barker, W.C., George, D.G., Mewes, H.W., Pfeiffer, F. and Tsugita, A. (1993) Nucleic Acids Res., 21 (13), 3089–3092.
14. Pattabiraman, N., Namboodiri, K., Lowrey, A. and Gaber, B.P. (1990) Protein Seq. Data Anal., 3, 387–405.
15. Warren Gish, NCBI (gish@ncbi.nlm.nih.gov). Software available via anonymous ftp on ncbi.nlm.nih.gov in the directory pub/nrdb.
16. Pappin, D.J.C., Horjup, P. and Bleasby, A.J. (1993) Current Biology, 3, 327–332.
17. Attwood, T.K. and Beck, M.E. (1994) Protein Engineering, 7 (7), 841–848.
18. Parry-Smith, D.J. and Attwood, T.K. (1992) CABIOS, 8 (5), 451–459.
19. Attwood, T.K. and Findlay, J.B.C. (1993) Protein Engineering, 6 (2), 167–176.
20. Parry-Smith, D.J. and Attwood, T.K. (1991) CABIOS, 7 (2), 233–235.
21. Attwood, T.K. and Findlay, J.B.C. (1994) Protein Engineering, 7 (2), 195–203.
22. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) J.Mol.Biol., 215 (3), 403–410.